

# PREDICTING CUSTOMER SHOPPING TRENDS USING RANDOM FOREST AND LOGISTIC REGRESSION TECHNIQUES

**Adam Abdurrahman, Gilang Ruen Soekarno, Aqeel Fazlemawla Fasliah, E.Faris Farhan Syah**

Faculty of Computer Science, Information System, President University, Jl. Ki Hajar Dewantara, Kota Jababeka, Cikarang Baru Bekasi 17550 - Indonesia

---

## Abstract

Customer shopping behavior is crucial for organizations when it comes to efficient implementation of marketing strategies and customer loyalty. The purpose of this research was to predict the buying behaviors of customers by categorizing the customers according to their buying behaviors. Logistic regression and Random Forest classification was used on sample data from Kaggle with variables such as age, purchase amount, product category, and number of purchases. It is important to note that the project was carried out using Python and Orange data analysis tools to prefix customer types, for instance 'frequent buyer' and 'discount hunter'. The results showed that the proposed model observed the classification accuracy standard for targeted marketing. Cross validation was used in evaluating the model and success was measured based on accuracy and validity across different customers segments. This project shows how great benefit could be accorded to strategic thinking in customer relations and marketing through use of data fragmented segmentation.

**Keywords:** Customer Shopping, Random Forest, Logistic Regression, Data Analyst, Orange

---

## I. INTRODUCTION

The development of digital technology has driven the rise of e-commerce, making transactions easier, faster and more efficient. Along with that, consumer behavior data continues to grow and grow. For businesses, understanding consumer spending patterns through data analysis is becoming increasingly important to make the right decisions. By utilizing this data, companies can develop more effective marketing strategies, increase sales, and strengthen customer loyalty.

To predict consumer behavior, Random Forest and Logistic Regression were chosen because both are able to provide accurate results and are easy to interpret. Random Forest simplifies complex data into a clear and understandable model, helping to identify key factors that influence purchases. Logistic Regression is used to predict the probability of consumer behavior based on

important variables. This method was chosen for its flexibility in handling large datasets and its ability to generate insights that can be implemented directly in business strategies.

With the application of these predictive methods, businesses can optimize operational strategies, adjust inventory, and anticipate changes in market demand. Ultimately, the use of Random Forest and Logistic Regression in predicting shopping trends not only helps improve marketing effectiveness but also provides a competitive advantage for companies amidst changing market dynamics.

### A. Previous Research

Intelligent Personalized Shopping Recommendation Using Clustering and Supervised Machine Learning Algorithms. This research employs characteristic analysis of grocery purchase patterns to develop a system that will be used for the generation of shopping recommendations using artificial

neural networks. It segments its customers using Random Forests as well as clustering and thus enables businesses to come up with differential incentives. This study shows that utilizing data for segmentation enhances the effectiveness of marketing that was also the aim for this project which seeks to implement the machine learning classification. (Chabane et al., 2022)

**LRFS: Online Shoppers' Behavior-Based Efficient Customer Segmentation Model.** This research builds an internet shopping behavior segmentation model, which concentrates on customer involvement and loyalty. Using both supervised learning, it categorizes online shoppers into different behavior-based groups to offer insights on enhanced customer interaction. The fact that this model deals with behavior-based segmentation makes it a good fit with the logistic regression and Random Forests that were used in this study to classify shopping behaviors. (Hayat Khan et al., 2024)

**An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market.** To cluster customers in the UK retail market, this paper reviews several clustering algorithms including K-means and Gaussian Mixture Models. It underlines how a data analytical approach to marketing can be useful when segmenting customers according to their shopping behaviors. It is mainly centered on clustering; nonetheless, it delivers a basis for the comprehension of customer segmentation which is applicable to this project's strategy of distal and proximal supervised learning techniques as employed in predictive modeling. (John et al., 2023)

## **B. Research Gap**

This research fills a gap in current empirical models of customer segmentation by shifting from broad, generic segments to the actionable, specific segments such as, Frequent Buyers, Discount hunters, and that are so essential for successful targeted marketing campaigns but understudied in prior research. Unlike other studies, which used clustering like K-means, or many

algorithms mentioned in the MDPI research, this study uses Random Forests and logistic

regression as supervised models to classify customers into the behavioral groupings. Customer motivations are well captured based on certain features such as the frequency of purchase, use of discounts and finally, buying seasons. Furthermore, the research uses Orange software and Python, underlining the practice possibilities based on currently available tools within organizations that may not possess a rich data science background. It enables businesses to adapt segmentation knowledge for practical segmentation marketing, customer interactions, and retention initiatives, thereby offering a practical alignment between/tools segmentation and its practical application.

## **C. Objective**

Customer shopping behavior is crucial for organizations when it comes to efficient implementation of marketing strategies and customer loyalty. The purpose of this research was to predict the buying behaviors of customers by categorizing the customers according to their buying behaviors. Logistic regression and Random Forest classification was used on sample data from Orange software and Kaggle with variables such as age, purchase amount, product category, and number of purchases. It is important to note that the project was carried out using Orange data analysis tools to prefix customer types, for instance 'frequent buyer' and 'discount seeker'. The results showed that the proposed model observed the classification accuracy standard for targeted marketing. Cross validation was used in evaluating the model and success was measured based on accuracy and validity across different customers segments. This project shows how great benefit could be accorded to strategic thinking in customer relations and marketing through use of data fragmented segmentation.

## II. METHOD

### A. Data Collection

In the *data collection* stage, we utilized a dataset obtained from the Kaggle platform. The dataset was then imported into the Orange software for further processing and analysis. Kaggle was selected as the data source due to the availability of relevant and comprehensive datasets to support the analysis of consumer behavior. The dataset is well-structured and it has no missing values, also most of the variables are categorical making them suitable for encoding techniques. The dataset has 3900 rows and 18 columns in which include:

- **Customer ID:** Unique identifier for each customer.
- **Age:** Age of the customer (integer).
- **Gender:** Categorical (Male/Female).
- **Item Purchased:** Categorical, e.g., Blouse, Pants, etc.
- **Category:** Categorical, e.g., Clothing, Accessories.
- **Purchase Amount (USD):** Numerical (range: 20-100 USD).
- **Location:** Categorical with 50 unique states.
- **Size:** Categorical (e.g., S, M, L, XL).
- **Color:** Categorical with 25 unique colors.
- **Season:** Categorical (e.g., Spring, Summer, Fall, Winter).
- **Review Rating:** Numerical (range: 2.5-5.0).
- **Subscription Status:** Categorical (Yes/No).
- **Shipping Type:** Categorical (e.g., Free Shipping, Standard).
- **Discount Applied:** Binary categorical (Yes/No).
- **Promo Code Used:** Binary categorical (Yes/No).

- **Previous Purchases:** Numerical (range: 1-50).
- **Payment Method:** Categorical (e.g., PayPal, Credit Card).
- **Frequency of Purchases:** Categorical (e.g., Monthly, Weekly).

*Frequency of Purchases* and *Discount Applied* serve as key variables in this study

### B. Data Pre Processing



Figure 1. Orange Configuration

The *data pre-processing* stage ensures that the dataset is prepared and optimized for analysis and model development. Two separate workflows were created with similar processing steps but different targets. The first workflow focuses on predicting *Frequency of Purchase* as the target variable, while the second targets *Discounts*.

Based on figure 1, the pre-processing steps in Orange involved the following:

1. **Edit Domain:** This node was used to define the target variable for each workflow. For example, *Frequency of Purchase* and *Discounts* were set as target variables in their respective workflows. Additionally, the data types of features were adjusted to align with the analysis requirements.
2. **Data Sampler:** This node split the dataset into two parts:
  - **Training data:** Used to train the machine learning models.

- **Test data:** Used to evaluate the performance of the trained models.

Both workflows utilized the same features, including numeric attributes like *Customer ID* and other supporting variables. The main goal of this stage was to ensure that the dataset was clean, structured, and ready for subsequent analysis.

While in Python, Target variable mapping was done first, where, for the target variable "Frequent Buyer", the Frequency of Purchase column was taken and preprocessed. Accordingly, 'Weekly', 'Bi-Weekly' and 'Fortnightly' purchase frequencies were encoded to 1, representing frequent buyers, while other frequencies, which are 'Monthly', 'Quarterly', 'Every 3 Months' and 'Annually', were encoded to 0, representing non-frequent buyers.

The target variable was obtained from the "Discount Applied" column, which took "Yes" as 1 for customers who got to use the discount and "No" as 0 otherwise. The categorical features are gender, category, location, size, color, season, subscription status, shipping type, and payment method; these have then been encoded using LabelEncoder into a numerical format. This will ensure feature compatibility for machine learning algorithms without affecting the importance of such variables in predictive modeling.

**Encoded Purchases of Frequency**

Weekly	Bi- Weekly	Fortnightly	Monthly	Quarterly	Every 3 Months	Annually
1	1	1	0	0	0	0

Table 1. Encoded Purchases of Frequency

**Encoded Discount Applied**

Yes	No
1	0

Table 2. Encoded Discount Applied

Some important features were filtered out and prepared for analysis: numeric features like Age, Purchase Amount in USD, Review Rating, Previous Purchases, among others, plus encoded categorical columns. It indicates features which are impactful in terms of customers' behavior. Based on the feature selection result, the dataset was divided into two training and testing subsets. Then, the function called the `train_test_split` is invoked with 80% assigned for training and the remaining 20% used for the purpose of testing. This model also uses a fixed `random_state` to ensure that its result is reproducible. More importantly, numerical features - Purchase Amount (USD) and Age were then scaled to improve the performance or optimization of Logistic Regression.

### C. DATA ANALYST

The *data analysis* phase involved the application of two machine learning algorithms:

1. **Logistic Regression:** A model used to analyze the relationship between independent variables and a categorical target variable. This is particularly useful for binary or multi-class classification problems.
2. **Random Forest:** An ensemble learning method that builds multiple decision trees to produce more accurate and robust predictions.

These algorithms were applied to the dataset to uncover patterns and relationships among variables. The analysis was conducted from two perspectives: one focused on *Frequency of Purchase* and the other on *Discounts*. This dual approach provided a more comprehensive understanding of consumer behavior.

Logistic Regression was used as a baseline to understand the relationship between independent variables with binary target variables of *Frequent\_Buyer* and *Discount\_Seeker*. To avoid class imbalances, `class_weight='balanced'` was used in training the models, making sure the minority classes were not ignored in the process.

Random Forest was another ensemble learning approach that helped in identifying the non-linear relationships and complex interactions within the data. A suite of decision trees is used in the model for better prediction accuracy and to make the model more robust. For training and testing the model, the same division as in Logistic Regression has been performed to keep the comparisons fair. Feature importance analysis has been done to identify the variables that most strongly contribute to the predictions, hence providing insights into the most important drivers of customer behavior. Further on, both models were compared against each other in terms of how effective they are in predicting the two types of buyers showcasing their respective strengths and weaknesses.

#### D. MODEL EVALUATION

The model evaluation phase aimed to assess the performance of each algorithm using relevant metrics. The following nodes were employed in Orange:

1. **Test and Score:** This node evaluated the models using the test dataset. It calculated metrics such as accuracy, AUC, precision, recall, and F1-score, which are critical for comparing the performance of Logistic Regression and Random Forest.
2. **Confusion Matrix:** This node provided detailed insights into the models' predictions, including the number of true positives, true negatives, false positives, and false negatives. This information is essential for identifying and addressing model errors.
3. **Predictions:** This node displayed the models' predictions on the test dataset, offering further insights into how the models performed on unseen data.

The evaluation process provided a clear understanding of each model's strengths and weaknesses in predicting the target variables (*Frequency of Purchase* and *Discounts*). The insights gained from these evaluations guided the selection of the most appropriate model for this analysis.

### III. RESULTS AND DISCUSSION

#### A. Result

##### 1. Frequent Buyer

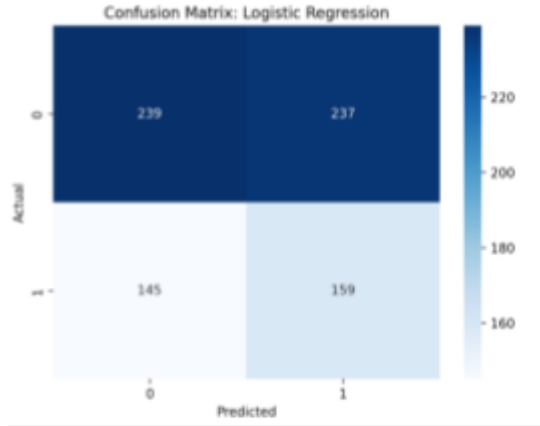


Figure 2. Frequent Buyer Confusion Matrix: Logistic Regression

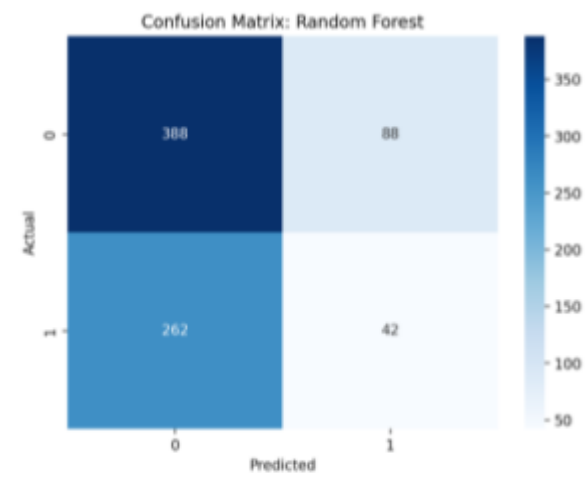


Figure 3. Frequent Buyer Confusion Matrix: Random Forest

Based on Figure 2 and Figure 3, the result of confusion matrices for the Logistic Regression and Random Forest models provide insights into their performance in predicting Frequent Buyers. In Logistic Regression, the model correctly predicted 239 non-frequent buyers as non-frequent buyers, which are the True Negatives, and 159 frequent buyers as frequent buyers, which are the True Positives. However, it also mislabeled 237 non-frequent buyers as

frequent buyers and failed to correctly identify 145 actual frequent buyers, misclassifying them as non-frequent buyers. This indicates that while Logistic Regression maintains a balance between predicting frequent and non-frequent buyers, it produces a significant number of false positives and false negatives, resulting in lower precision and recall. The model exhibits a slight bias toward predicting customers as frequent buyers, even when they are not, which limits its accuracy.

Conversely, the Random Forest model strongly leans toward classifying non-frequent buyers. It correctly predicted 388 actual non-frequent buyers as non-frequent buyers (True Negatives) while incorrectly classifying only 88 non-frequent buyers as frequent buyers (False Positives). However, its recall performance is poor, as it correctly identified only 42 actual frequent buyers (True Positives) while failing to classify 262 frequent buyers, mislabeling them as non-frequent buyers (False Negatives).

Frequent Buyer Prediction Results					
Model Evaluation					
	Accuracy	AUC	F1 Score	Precision	Recall
Logistic Regression	51.83%	51.30%	48.43%	48.13%	52.38%
Random Forest	58.13%	58.12%	18.33%	32.33%	13.83%

Figure 4. Frequent Buyer Model Evaluation

Based on Figure 4, this bias in predicting non-frequent buyers results in high precision but very low recall, indicating the model's difficulty in capturing most actual frequent buyers. In comparison, Logistic Regression achieves better recall, capturing 159 true frequent buyers compared to Random Forest's 42. However, Random Forest outperforms in precision, with fewer false positives (88) than Logistic Regression (237). Overall, both models struggle to effectively predict frequent buyers, likely due to insufficient or noisy features in the dataset. While Logistic Regression offers a more balanced prediction, Random Forest's

precision comes at the cost of missing a significant number of frequent buyers.

Logistic Regression achieves a better balance between recall and precision, slightly better at identifying frequent buyers but struggles with false positives, leading to moderate precision. While Random Forest performs well at correctly identifying non-frequent buyers, very poor recall for frequent buyers which mean Random Forest misses most of them. In conclusion, Logistic Regression is more suitable if the goal is to identify as many Frequent Buyers as possible, even if it means tolerating some false positives. Meanwhile, Random Forest is more conservative, better at avoiding false positives but very poor at capturing frequent buyers.

## 2. Discount Seeker

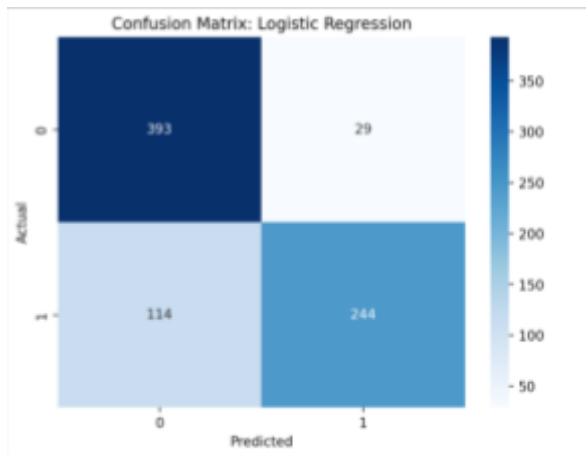


Figure 5. Discount Seeker Confusion Matrix: Logistic Regression

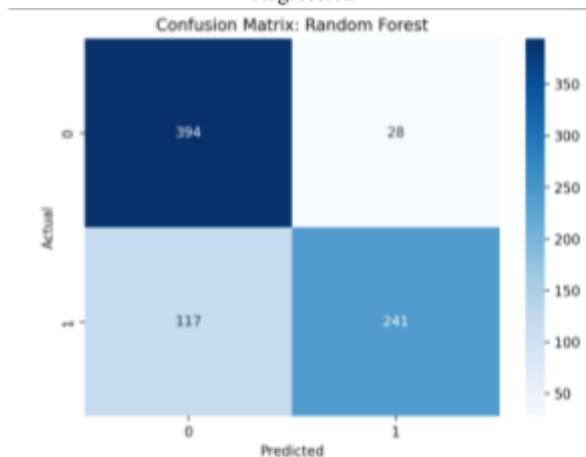


Figure 6. Discount Seeker Confusion Matrix: Random Forest

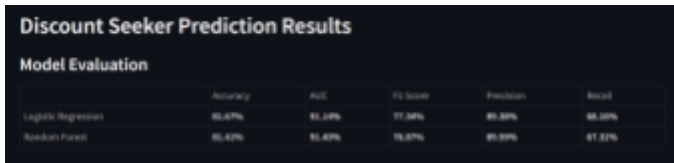
Based on Figure 5 and Figure 6, the result of confusion matrices for the Logistic Regression and Random Forest models reveal their effectiveness in predicting Discount Seekers. These matrices showcase the models' ability to classify individuals into two categories: non-discount seekers (0) and discount seekers (1).

For Logistic Regression, the model successfully identified 393 non-discount seekers as non-discount seekers (True Negatives) and 244 discount seekers as discount seekers (True Positives). It misclassified 29 non-discount seekers as discount seekers (False Positives) and missed 114 actual discount seekers, incorrectly labeling them as non-discount seekers (False Negatives). This performance indicates that Logistic Regression is effective at distinguishing between the two classes, achieving a balance between precision and recall with a high rate of accurate predictions and relatively few misclassifications.

The Random Forest model demonstrated slightly different results. It correctly identified 394 non-discount seekers (True Negatives), a marginal improvement over Logistic Regression, but identified fewer discount seekers, with 241 classified correctly (True Positives). The model misclassified 28 non-discount seekers as discount seekers (False Positives) and missed 117 discount seekers (False Negatives). These results suggest that while Random Forest minimizes false positives slightly better than Logistic Regression, it sacrifices a small degree of recall by correctly identifying fewer discount seekers.

Both models excel in identifying non-discount seekers, as evidenced by their high True Negative counts and low False

Positive rates. However, they encounter challenges in accurately capturing all discount seekers, reflected in their False Negative counts. Logistic Regression demonstrates slightly better recall, identifying more true discount seekers (244 compared to 241 for Random Forest). On the other hand, Random Forest achieves better precision by reducing false positives (28 compared to 29 for Logistic Regression).



Model Evaluation					
	Accuracy	AUC	F1 Score	Precision	Recall
Logistic Regression	85.67%	81.34%	77.34%	80.30%	86.50%
Random Forest	85.42%	81.40%	76.87%	81.00%	81.33%

Figure 5. Discount Seeker Model Evaluation

Based on Figure 6, Logistic Regression, for its part, has higher recall than the Random Forest by a small margin because it was better at predicting the true positives of the Discount Seekers. The precision for this classifier is competitive because most of the predicted classes of being Discount Seekers are indeed correct, though a bit low compared to that of Random Forest. Overall, Logistic Regression is balanced with an F1 score of 77.34%. On the other hand, Random Forest provides a slightly better precision compared with Logistic Regression, hence less number of false positives while predicting Discount Seekers. The recall is slightly lower, with a slightly reduced F1 score at 76.87%, which means the performance concerning precision and recall is not that well-balanced. Therefore, Logistic Regression is relatively better in identifying Discount Seekers, with high recall and balanced metrics. Random Forest is somewhat more conservative, strong at precision, and better suitable when avoidance of false positives is in priority, while the cost of missing more potentially sought-after Discount Seekers.

In summary, both models are effective tools for predicting discount seekers, with minor differences in their trade-offs between precision and recall. Logistic Regression slightly favors recall, making it better at identifying more discount seekers, while Random Forest offers improved precision, reducing false alarms. The choice between these models depends on the business goal—whether maximizing recall or precision takes priority. Both are viable for deployment based on these requirements.



## B. Discussion

The comparison between Logistic Regression and Random Forest models in our analysis revealed distinct trade-offs in precision and recall. Logistic Regression demonstrated balanced performance, albeit with notable false positives and negatives, aligning with insights from (Hayat Khan et al., 2024), who emphasized the challenges of achieving granularity in customer segmentation. Their LRFS model addressed these issues by incorporating a novel "S" (Staying Rate) component to enhance precision and recall through dimensionality reduction techniques like PCA and t-SNE(LRFS\_Online\_Shoppers\_Be...). Our model could benefit from a similar approach to improve its ability to classify frequent buyers accurately.

Similarly, the UK retail segmentation study by (John et al., 2023) showed the effectiveness of Gaussian Mixture Models (GMM) over Random Forest for better customer segmentation precision. This indicates that Random Forest's tendency to favor non-frequent buyer classification, as seen in Our results, might be mitigated by exploring probabilistic clustering approaches(analytics-02-00042).

The analysis of Discount Seekers highlighted Logistic Regression's higher recall and Random Forest's superior precision. These results resonate with (Chabane et al., 2022), who explored recommender systems using clustering and supervised learning to balance recall and precision. They demonstrated that Random Forest, while strong in precise classifications, often lags in capturing broader patterns due to its deterministic nature(journal.pone.0278364). Logistic Regression's balanced approach in Our study

could mirror the benefits of a hybrid recommender system that combines collaborative and content-based filtering for more effective segmentation.

Furthermore, the LRFS study's use of dimensionality reduction to optimize clustering accuracy can be adapted for Discount Seekers. By integrating unsupervised techniques such as k-means or autoencoders, Our segmentation models might better handle the nuances of customer behavior across price-sensitive segments(LRFS\_Online\_Shoppers\_Be...)(journal.pone.0278364).

The analysis of our segmentation models suggests several avenues for improvement to align with insights from prior research. Dimensionality reduction techniques, such as PCA or t-SNE, could be utilized to refine the feature space for both Frequent Buyers and Discount Seekers, enhancing the models' ability to capture nuanced patterns in customer behavior. Furthermore, the integration of probabilistic clustering methods like Gaussian Mixture Models (GMM) for Frequent Buyers and hybrid recommender systems for Discount Seekers can address trade-offs in precision and recall by leveraging probabilistic insights and combining collaborative and content-based filtering. Additionally, incorporating advanced feature engineering metrics, such as the "Staying Rate" or price sensitivity, can enhance model interpretability and segmentation accuracy, as demonstrated in the LRFS model and related studies. These approaches collectively offer a pathway to improve segmentation outcomes, aligning the models more closely with business objectives and customer-centric strategies.

By incorporating insights from the referenced research, our segmentation models can be refined to provide actionable results, better aligning with strategic marketing goals and customer-centric approaches.

#### IV. CONCLUSION

The goal of this study was to analyze customer shopping behaviors and classify them into actionable segments, such as Frequent Buyers and Discount Seekers, using Logistic Regression and Random Forest models. Logistic Regression was used as the baseline model due to its simplicity and interpretability, and it demonstrated slightly better recall, especially in identifying Discount Seekers. Random Forest, in contrast, excelled in precision, minimizing false positives but with lower recall, particularly for Frequent Buyers. Both models effectively identified non-discount seekers and non-frequent buyers, although predicting frequent buyers remained challenging. Random Forest also showed a tendency to favor non-frequent buyers, indicating some bias.

Overall, the study successfully achieved its objectives, demonstrating the

value of these models in supporting strategic marketing decisions, customer loyalty programs, and resource allocation for better business outcomes.

#### REFERENCES

- Chabane, N., Bouaoune, A., Tighilt, R., Abdar, M., Boc, A., Lord, E., Tahiri, N., Mazouze, B., Rajendra Acharya, U., & Makarenkov, V. (2022). Intelligent personalized shopping recommendation using clustering and supervised machine learning algorithms. *PLoS ONE*, 17(12 December).  
<https://doi.org/10.1371/journal.pone.0278364>
- Hayat Khan, R., Fabian Dofadar, D., Alam, M. G. R., Siraj, M., Rafiul Hassan, M., & Mehedi Hassan, M. (2024). LRFS: Online Shoppers' Behavior-Based Efficient Customer Segmentation Model. *IEEE Access*, 12, 96462–96480.  
<https://doi.org/10.1109/ACCESS.2024.3420221>
- John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, 2(4), 809–823.  
<https://doi.org/10.3390/analytics2040042>